Test-enhanced learning of clinical reasoning: a cross-over randomised trial

Tobias Raupach, MD, MME[1,2]: raupach@med.uni-goettingen.de
Jil Cathérine Andresen[1]: jil.andresen@googlemail.com
Katharina Meyer[1]: katharina.meyer@med.uni-goettingen.de
Lisa Strobel, MD[1]: strobell@ymail.com
Michael Koziolek, MD[3]: mkoziolek@med.uni-goettingen.de
Wolfram Jung, MD[4]: wolfram.jung@med.uni-goettingen.de
Jamie Brown, PhD[2]: jamie.brown@ucl.ac.uk
Sven Anders, MD[5]: s.anders@uke.de

[1] Clinic for Cardiology and Pneumology, Göttingen University Medical Centre, Robert-Koch-Straße 40, Göttingen, D-37075, Germany

[2] Health Behaviour Research Centre, University College London, 1-19 Torrington Place, London, WC1E 7HB, United Kingdom

[3] Clinic for Nephrology and Rheumatology, Göttingen University Medical Centre, Robert-Koch-Straße 40, Göttingen, D-37075, Germany

[4] Clinic for Haematology and Oncology, Göttingen University Medical Centre, Robert-Koch-Straße 40, Göttingen, D-37075, Germany

[5] Department of Legal Medicine, University Medical Centre Hamburg-Eppendorf, Butenfeld 34, Hamburg, D-22529, Germany

Address for Correspondence:

Prof. Tobias Raupach, MD, MME

University Hospital Göttingen

Department of Cardiology and Pneumology

Robert-Koch-Straße 40

D-37075 Göttingen, Germany

Phone: +49 551 39-8922

Fax: +49 551 39-20900

E-mail: raupach@med.uni-goettingen.de

word count (excluding abstract, tables and references): 3687

Abstract

Context: Clinical reasoning is an essential skill, the foundations of which need to be acquired during undergraduate medical education. Student performance with regard to clinical reasoning can be assessed using key feature examinations. However, within a paradigm of test-enhanced learning, such examinations may also be used to enhance long-term retention of procedural knowledge relevant to clinical reasoning.

Objectives: This study tested the hypothesis that repeated testing with key feature questions is more effective than repeated case-based learning in fostering clinical reasoning.

Methods: In this randomised cross-over trial, fourth-year undergraduate medical students attended ten weekly computer-based seminars during which patient case histories covering general medical conditions were displayed. Presentation format was switched between groups every week: In the control condition, students studied long case narratives. The same content was covered in the intervention condition, but here case presentation was augmented by a sequence of key feature questions. Using a within-subjects design, student performance on intervention and control items was assessed thirteen weeks (exit exam) and nine months (retention test) after the first day of term .

Results: A total of 87 out of 124 eligible students provided complete data for the longitudinal analysis (response rate 70.2%). In the retention test, mean ± standard deviation student scores on intervention items were significantly higher than on control items (56.0 ± 25.8% vs. 48.8 ± 24.7%, $p < 0.001$). The results remained unchanged after accounting for exposure time in a linear regression analysis that also adjusted for sex and general student performance levels.

Conclusions: This is the first study to demonstrate an effect of test-enhanced learning on clinical reasoning as assessed with key feature questions. In this randomised trial, repeated testing was more effective than repeated case-based learning alone. Curricular implementation of longitudinal key feature testing may considerably enhance student learning outcome on relevant aspects of clinical medicine.

Key words: medical education, assessment, learning, testing effect, retention, clinical reasoning

Introduction

The overall aim of undergraduate medical education is to equip future physicians with a solid knowledge base and essential skills required to manage their future patients. Amongst the higher-order cognitive functions that need to be mastered is applying knowledge to new clinical situations in order to make informed choices about diagnostic procedures and therapeutic options. These functions are usually described as clinical reasoning or clinical cognition, and they encompass a range of strategies used to generate and test hypotheses about a disease or mechanism underlying a given patient's presenting complaint as well as to arrive at decisions based on judgements of the prognostic significance of diagnostic tests [1]. At least two processes are involved in clinical reasoning, the first one being intuitive (e.g., pattern recognition from first impressions) and the second one being analytic in nature [2]. It has been proposed that easier cases can readily be solved by applying intuitive strategies while more in-depth processing is needed when dealing with more complex cases [3]. Clinical reasoning hinges on the availability of a solid knowledge base, deliberate practice and metacognitive processes (e.g., reflective coping [4]).

The most popular teaching formats aimed at fostering clinical reasoning are small-group sessions (e.g. in problem-based or case-based learning [5, 6]) and clinical rotations. However, since clinical teachers are required to facilitate small-group teaching, this instructional format is resource-intensive. In addition, the larger group format is not well-suited to students who learn at different speeds and is prone to particular students dominating discussions, thereby possibly hampering learning outcome in others [7]. These problems should not occur during clinical attachments where one-to-one teaching is available from experienced clinicians [8]. However, the content for this format cannot be standardised to the same extent as in formal teaching sessions. Learning opportunities on medical wards depend on the type of patients treated on these wards. As a consequence, students might be less likely to encounter pa-

tients with specific diseases (including some for which swift decision-making is crucial), thus limiting the range of content covered during clinical attachments.

One way to standardise the content covered and provide equal learning opportunities to all students is to use computer-assisted case-based learning [9]. Despite being more effective than no teaching at all, computer-based learning is in itself not generally superior to more traditional instructional formats [10]. In order to be effective, computer-based teaching interventions should be informed by theory. One recent advancement in educational psychology refers to test-enhanced learning. While it has been known for some time that summative assessments drive student learning *behaviour* [11], the first studies assessing the impact of formative assessments on learning *processes* in medicine were only published in 2008 [12] (see [13] for a thorough discussion of the principles underlying test-enhanced learning). These trials tended to focus on rote memory [14] rather than complex cognitive functions, and some included short follow-up periods (e.g., one week [15]). While some studies used sophisticated testing methods [16], others relied on multiple choice questions [17] which have been shown to be less effective in this scenario [13] and potentially even detrimental to student learning outcome [18]. Current research on test-enhanced learning addresses mechanisms underlying the observed effects. For example, superior performance in a final test could either be the result of effortful retrieval during preceding tests ('depth of processing hypothesis' [19]) or it might simply reflect a training effect if the format of these tests is similar, thus evoking identical cognitive processes ('transfer appropriate hypothesis' [20]). In a recent study comparing these hypotheses, McConnell et al. [21] found that student performance was better when more demanding test formats such as short answer questions and context-rich multiple choice questions (as opposed to context-free questions) were used during the training phase. However, the effect seen in the final test was limited to questions that were identical to those seen during the training. From a practical perspective, medical educators need to know how repeated testing can be used effectively and efficiently at the same time.

Given the potential lack of exposure to some relevant aspects of clinical medicine during small-group and face-to-face teaching as well as the opportunities arising from the concept of test-enhanced learning, the present trial combined computer-based learning with key feature questions to assess whether repeated testing is more effective than repeated case-based learning in fostering clinical reasoning [22]. We hypothesised that undergraduate medical students answering key feature questions would retain more procedural knowledge over a six-month period than students being exposed to the same content but without being prompted to answer key feature questions. The effect – if present – was expected to be attributable to enhanced ability to access the relevant knowledge rather than a mere training effect regarding the handling of key feature questions.

Methods

Study design

This was a randomised, controlled, non-blinded cross-over trial involving fourth-year undergraduate medical students. At the institution where the study was conducted, undergraduate medical education is divided in two parts: During the first two years, basic sciences as well as anatomy, biochemistry, physiology and medical psychology are taught. After passing a high-stakes exam, students progress to the clinical phase of the curriculum. This phase is made up of 21 modules, each lasting between two and seven weeks. Modules in the first year of the second phase are devoted to practical skills training, microbiology, pharmacology and radiology, while modules in the second and third year focus on signs and symptoms of specific diseases and approaches to treatment. In this study, fourth-year students enrolled in three consecutive teaching modules (cardiology/pulmonology, nephrology/rheumatology and haematology/oncology) in winter term 2013/14 were eligible for study participation. Following an entry exam in the first week of term, students were stratified by sex and performance levels and randomised to one of two groups. Students in both groups attended weekly comput-

er-assisted learning sessions ('e-seminars'; maximum duration 45 minutes) during which they were presented with four patient case histories related to teaching content of the previous week (see **Table 1**). Case content was identical for all students but the format varied each week between 'studying only cases' or 'testing cases', in that the latter also involved prompts to answer five key feature questions per case. These questions are referred to as 'items', and all analyses were done on the item level (and not on the level of complete cases). Group A began with testing cases (see **Figure 1**: black boxes), and Group B began with studying only cases (grey boxes). Presentation format then alternated systematically for each group across ten sessions. Studying only cases yielded exactly the same information as testing cases, the only difference being the presence of five questions ('items') each in the latter. Thus, each student answered a total of 100 items during the five intervention e-seminars he/she was assigned to. Of these, 15 occurred in two e-seminars, allowing for repeated testing of these 'intervention items'. Since content was mutually exclusive for the two groups, intervention items for group A were never presented as questions in group B (and vice versa). Thus, by default, intervention items in group A served as control items in group B. Any student – regardless of group assignment – was exposed to 15 intervention items in two e-seminars and to 15 control items in two different e-seminars. The content of these 30 items had been pre-specified based on item characteristics obtained in a pilot phase in the preceding term. They were compiled into four cases containing 6 or 9 key features each (with random allocation of items to cases), and the diseases covered are presented in **Table 2**. In order to ensure comparability, the same cases were used in the entry and exit exam as well as the retention test, but they were different from the cases presented in e-seminars occurring between the entry and exit exam. All exams were formative in nature, i.e. no incentive to achieve high scores was provided as this may have confounded results [11].

Detailed feedback was created for each item, and students were able to access this feedback by clicking on a specific button. The same information provided in the feedback to specific questions was also provided as 'background information' in the control condition. Feed-

back for intervention items contained all options from the key feature long menu that were considered ‘correct’. It also contained explanations on why some distractor options anticipated to be chosen by students were deemed ‘incorrect’. The presentation of the information differed between groups in that feedback for intervention items was tailored to the question while all relevant information was included in the ‘background information’ without highlighting those aspects that were tested in the intervention group. Content covered in e-seminars was not repeated during formal teaching, thus any increase in knowledge from the first intervention e-seminar to the exit exam would be due to exposure to the content during the e-seminar or self-study.

Student enrolment, data collection and analysis

Four weeks before the start of winter term 2013/14, students were informed about the study by e-mail. On the first day of term, they were invited to provide written consent to participate in the study, and consenting students were followed up over a total of nine months. All three examinations as well as all ten e-seminars were held in the institution’s e-learning resource centre. Each student was assigned to one of the two computer rooms at the resource centre, and presentation format (testing cases / studying only cases) was switched for each room on a weekly basis. Students not reporting to the same room for every e-seminar were excluded from the analysis as they received a higher or lower dose of the intervention, thus contaminating results. Exams and e-seminars were scheduled to last no longer than 45 minutes within which students were free to complete sessions at their own pace. Individual login time was recorded.

The primary outcome for this study was the within-subject difference in percent scores in intervention versus control items in the retention test six months after the last e-seminar (paired t test). Pearson correlations were calculated to assess the moderating effect of exposure time to studying/testing material on retention test performance. In addition, linear re-

gression analyses were run using the difference between scores in testing and studying items in the exit and retention test as the dependent variables and the difference between mean time spent on testing and studying only cases as the independent variable, adjusting for student sex and percent scores in summative end-of-module examinations as a surrogate parameter for overall performance levels. Secondary outcomes included student performance in the exit exam and the proportion of correct answers on individual items in both exams.

Statistical analysis was performed using SPSS 22.0 (SPSS Inc., Chicago, Illinois, USA). Data are presented as mean ± SD or percentages (n) unless otherwise stated. Significance levels were set to $p<0.05$. This study was approved by the local Ethics Committee (application number 2/10/13).

Results

Student characteristics

The flow of participants through the study is displayed in **Figure 1**. Of the 124 students eligible for study participation, four did not provide written consent. Following exclusion of students due to missing data or contamination (see above), complete data were available for 87 students (longitudinal sample; effective response rate 70.2%). The mean age of study participants was 25.0 ± 2.9 years, and 58.6% (n = 51) were female. There were no significant differences regarding age, sex distribution and previous exam scores between students in the longitudinal sample and the 33 students excluded from the analysis (data not shown).

Descriptive analysis of e-seminars and exams

Attendance at e-seminars was high with no student missing more than two seminars in total. Mean login time at e-seminars was 17:46 ± 04:44 minutes, and students spent significantly more time on testing items than on studying items (21:47 ± 05:27 mins vs. 13:30 ± 05:24 mins per seminar; $p < 0.001$). The percentage of points scored on testing items ranged from 38.6% to 65.6%, reflecting the absence of a floor or ceiling effect. Internal consistency of e-seminars (testing items) tended to increase over the course of the study with the final four sets of testing seminars (20 items each) all yielding Cronbach's α values of >0.8. Cronbach's α of the entry, exit and retention exam was 0.663, 0.905 and 0.895, respectively.

Learning outcome

Mean percent scores in the entry, exit and retention exams were 22.6 ± 11.3%, 53.0 ± 24.4% and 52.4 ± 23.4%, respectively. **Figure 2** shows superior performance on testing compared with studying items by the primary measure of long-term retention. Percent scores for testing items and studying items were 56.0 ± 25.8% and 48.8 ± 24.7%, respectively ($p < 0.001$). The difference between the two scores was slightly larger on the secondary measure of exit exam performance (59.3 ± 27.7% vs. 46.7 ± 24.8%, $p < 0.001$). There was a significant correlation between excess time spent on testing items and the difference between performance in testing and studying items in the retention test ($r = 0.273$; $p = 0.011$) but not in the exit exam ($r = 0.038$; $p = 0.724$). After adjusting for student sex and performance in summative end-of-course examinations, the main effect of the intervention on exam performance was still significant for both the retention test (intercept: -7.03; 95% CI -12.03-[-2.03]; $p = 0.006$) and the exit exam (intercept: -8.40; 95% CI -13.71-[-3.09]; $p = 0.002$).

Student performance on individual items in the exit and retention test is given as a function of prior exposure (repeated testing vs repeated study) in **Table 2**. Performance trajectories (**Figure 3**) indicate that once an intervention item had been answered correctly, another correct answer was provided on a subsequent test in 72% of cases. More importantly, 33% of

students who failed to provide a correct answer in one test went on to submit a correct answer in a subsequent test of the same item, thus suggesting a learning effect elicited by the feedback provided.

Discussion

To our knowledge, this is the first study using repeated key feature examinations in a test-enhanced learning paradigm. It is also the first study specifically addressing clinical reasoning in undergraduate medical students. After six months, we found a sustained effect of repeated testing that was robust even when accounting for additional exposure time to the material in testing sessions. The key feature examinations used in our study covered a wide range of diseases. Students who had been repeatedly tested outperformed students who had repeatedly studied the same material on tasks such as diagnosing pneumonia on a chest X-ray, detecting bronchial obstruction in a lung function test and calculating the Wells Score before ordering a D-dimer test in suspected pulmonary embolism, all of which are highly relevant to clinical practice. Once established, key feature testing appears to be a time-efficient way of enhancing retention of clinical reasoning. For some learning objectives, it may thus be an intriguing alternative to more resource-intensive teaching formats.

Comparison with previous research

Previous studies on the effectiveness of test-enhanced learning in medical education differ with regard to study design (observational [23] or randomised [24]), study population (medical [25], dental [15], and nursing [17] students), career level (from incoming students [23] to residents [26]), sample size (21 [23] to 138 [27]), learning objectives (from factual knowledge [14] to clinical applications [26]), testing format (from multiple choice questions [17] to essays [25] and simulated patient encounters [16]), number of interventions (one [14], three [26] or four [25]), and length of follow-up (between one week [17] and six months [16]). While this huge variation limits the extent to

which our results can be compared to those of earlier studies, there are some similarities between the present study and recent reports by Larsen and colleagues who studied test-enhanced learning in neurology. In one randomised cross-over study [26], 40 neurology residents who had received teaching on myasthenia gravis and status epilepticus subsequently participated in three testing or studying sessions over a period of four weeks. Six months later, performance on short-answer questions was better in items that had been repeatedly tested (percent score 39% vs. 26%). In another cross-over trial done by the same group [25], 47 first-year medical students received teaching on seizures, optic neuritis, myasthenia gravis and migraine. In a retention test taken six months later, repeated testing was associated with significantly higher percent scores than repeated study (40% vs. 20%). Similar results were obtained in a more recent trial involving sophisticated methods [16]: The authors used simulated patients for intervention sessions and the final retention test. While this approach helped demonstrate that the effects elicited by test-enhanced learning may translate into clinical practice, it also raises the issue of resource allocation within medical education [28]. Repeated testing with simulated patients is time-consuming and requires considerable staff numbers. Given the economies of scale associated with e-learning interventions [29] and the educational rationale for using key feature questions to foster complex cognitive functions [12], our study contributes to the growing body of literature on how e-learning can be used efficiently to improve student learning outcome [30]. Repeated key feature examinations can create learning opportunities that might otherwise not be available to all students (e.g., performing an arterial blood gas analysis in a case of suspected CO_2 intoxication, **Table 2**).

At first glance, the overall performance in the retention test with a mean score of just over 50% is disappointing. However, the percent scores achieved for intervention and control items compare favourably to earlier studies mentioned above [25, 26]. The results are likely due to the retention test being formative in nature. As these assessments did not create credit points, students did not revise for them, thus avoiding confounding by the impact of examination consequences on learning behaviour. Accordingly, our data likely reflect the true effect of

repeated testing on the ability to access the knowledge needed to solve the clinical problems presented.

One unexpected finding of the present study was the absence of a substantial decline in exam performance over a period of six months. Few studies on test-enhanced learning in medical education have specifically reported on retention but those that did suggest a considerable decrease in performance levels over the course of six months [14, 16, 25, 26]. As in most other studies [14, 17, 26], the exit exam and retention test were identical. Consequently, one potential explanation for the lack of a performance loss is that students might simply have remembered the cases, questions and answers. This is unlikely given that students were not reminded of the content over a period of six months. The magnitude of this effect would be expected to be similar for intervention and control items, and even if this effect did have an impact on our results, it was too weak to mask the sustained advantage of testing over studying. An alternative explanation for the excellent retention observed in this study is that students might have applied their knowledge in consecutive clinical attachments. This – and the fact that repeated study appeared to elicit superior retention in some items (see **Table 2**) – highlights the need for a multifaceted approach to teaching, blending traditional teaching formats with innovative concepts.

Strengths and limitations of the study

This study was designed according to current recommendations for this type of research [12]. We used a production test format rather than a recognition test format to assess clinical reasoning. Testing was spaced across weeks, and students received immediate educational feedback on their answers. The response rate was acceptable and there was no evidence of selection bias threatening the validity of the findings. Internal consistency of the exit and retention test was excellent. This study was not designed to address the cognitive mechanisms underlying the observed effect. However, our results are in line with the hypothesis that re-

peated testing with key feature questions enhances knowledge retention. Owing to the crossover design and the within-subjects comparison of intervention and control items, our data suggest that the observed score difference is due to an effect of repeated testing on the storage and retrieval of knowledge rather than an effect of training how to answer key feature questions. The latter is unlikely since all students received the same 'dose' of key feature training. Unlike another recent study [21], we found an overall effect of repeated testing despite the cases presented in the exit exam being considerably different from the cases presented during e-seminars, although identical learning objectives were targeted. We cannot comment on the cognitive processes underlying clinical reasoning that were primarily invoked by our intervention. It may be hypothesised that cases were easy enough to be solved by an intuitive strategy, thus requiring less higher-order metacognitive activity [1]; however, this hypothesis needs to be tested in future trials.

Despite these strengths, the generalisability of our findings is limited by the monocentric nature of the study. As this study was meant to shed light on the real-world effectiveness of test-enhanced learning, some potential confounding factors were not experimentally controlled. Most importantly, we did not collect data on independent student learning. Testing cases might have stimulated self-study outside the e-learning resource centre to a greater extent than reading cases, thus magnifying the observed effect. Unlike laboratory studies on test-enhanced learning, we did not attempt to control time-on-task during e-seminars but allowed students to terminate their sessions at their own convenience. However, login time was recorded and included in our analysis. There was weak evidence of an excess in time-on-task in testing sessions being associated with better learning outcome in testing items, but the effect of the intervention persisted when this was controlled for in the analysis. Finally, this study did not assess whether repeated testing with key feature questions impacts on student performance in the clinical setting. Although one study suggests such a link [16], more research is needed to establish a causal relation between repeated testing and more favourable patient outcomes.

Conclusions

For the first time, this study demonstrates that repeated case-based learning augmented with formative key feature questions is more effective than case-based learning alone in fostering clinical reasoning in undergraduate medical students. The effect was sustained over six months and cannot be explained by differences in exposure time between the two presentation formats. Repeated testing using key feature questions appears to be a time-efficient way of enhancing retention of procedural knowledge in undergraduate medical education.

Authors' contributions

TR conceived of the study, developed its design, drafted key feature cases, analysed the data and wrote the manuscript. JA facilitated e-seminars by uploading case material and was involved in data analysis. KM facilitated e-seminars and contributed to case upload. LS compiled studying only cases from testing cases and drafted feedback for all 200 key feature questions. MK drafted and revised key feature questions related to nephrology and rheumatology. WJ commented on the manuscript. JB was involved in data analysis and contributed to the Introduction and Discussion section. SA helped to design the study, provided advice on data presentation and commented on various versions of the manuscript.

Competing interests

None of the authors have any conflict of interest to declare.

Funding sources

This study was funded by a Fellowship for Innovations in University Teaching issued by the Stifterverband für die Deutsche Wissenschaft (project number H120 5228 5008 23472). JB is funded by a fellowship from the Study for the Society of Addiction.

Acknowledgements

We would like to thank all medical students who devoted their time to this study.

References

[1] Kassirer JP. Teaching clinical reasoning: case-based and coached. *Acad Med*. 2010;**85**:1118-1124.
[2] Croskerry P. A universal model of diagnostic reasoning. *Acad Med*. 2009;**84**:1022-1028.
[3] Elstein AS, Schwartz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*. 2002;**324**:729-732.
[4] Dunphy BC, Cantwell R, Bourke S, et al. Cognitive elements in clinical decision-making: toward a cognitive model for medical education and understanding clinical reasoning. *Adv Health Sci Educ Theory Pract*. 2010;**15**:229-250.
[5] Srinivasan M, Wilkes M, Stevenson F, Nguyen T, Slavin S. Comparing problem-based learning with case-based learning: effects of a major curricular shift at two institutions. *Acad Med*. 2007;**82**:74-82.
[6] Dequeker J, Jaspaert R. Teaching problem-solving and clinical reasoning: 20 years experience with video-supported small-group learning. *Med Educ*. 1998;**32**:384-389.
[7] Jaques D. Teaching small groups. *BMJ*. 2003;**326**:492-494.
[8] Gordon J. ABC of learning and teaching in medicine: one to one teaching and feedback. *BMJ*. 2003;**326**:543-545.
[9] Wahlgren CF, Edelbring S, Fors U, Hindbeck H, Stahle M. Evaluation of an interactive case simulation system in dermatology and venereology for medical students. *BMC Med Educ*. 2006;**6**:40.
[10] Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. *JAMA*. 2008;**300**:1181-1196.
[11] Raupach T, Brown J, Anders S, Hasenfuss G, Harendza S. Summative assessments are more powerful drivers of student learning than resource intensive teaching formats. *BMC Medicine*. 2013;**11**:61.
[12] Larsen DP, Butler AC, Roediger HL, 3rd. Test-enhanced learning in medical education. *Med Educ*. 2008;**42**:959-966.
[13] Roediger HL, Karpicke JD. The Power of Testing Memory - Basic Research and Implications for Educational Practice. *Perspect Psychol Sci*. 2006;**1**:181-210.
[14] Schmidmaier R, Ebersbach R, Schiller M, Hege I, Holzer M, Fischer MR. Using electronic flashcards to promote learning in medical students: retesting versus restudying. *Med Educ*. 2011;**45**:1101-1110.
[15] Baghdady M, Carnahan H, Lam EW, Woods NN. Test-enhanced learning and its effect on comprehension and diagnostic accuracy. *Med Educ*. 2014;**48**:181-188.
[16] Larsen DP, Butler AC, Lawson AL, Roediger HL, 3rd. The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Adv Health Sci Educ Theory Pract*. 2013;**18**:409-425.
[17] Dobson JL, Linderholm T. Self-testing promotes superior retention of anatomy and physiology information. *Adv Health Sci Educ Theory Pract*. 2014.
[18] Roediger HL, 3rd, Marsh EJ. The positive and negative consequences of multiple-choice testing. *J Exp Psychol Learn Mem Cogn*. 2005;**31**:1155-1159.
[19] Wheeler MA, Ewers M, Buonanno JF. Different rates of forgetting following study versus test trials. *Memory*. 2003;**11**:571-580.
[20] Morris CD, Bransford JD, Franks JJ. Levels of processing versus transfer appropriate processing. *J Verb Learn Verb Behav*. 1977;**16**:519-533.
[21] McConnell MM, St-Onge C, Young ME. The benefits of testing for learning on later performance. *Adv Health Sci Educ Theory Pract*. 2015;**20**:305-320.
[22] Karpicke JD, Roediger HL, 3rd. The critical importance of retrieval for learning. *Science*. 2008;**319**:966-968.

[23] Logan JM, Thompson AJ, Marshak DW. Testing to enhance retention in human anatomy. *Anat Sci Educ*. 2011;**4**:243-248.
[24] Kromann CB, Jensen ML, Ringsted C. The effect of testing on skills learning. *Med Educ*. 2009;**43**:21-27.
[25] Larsen DP, Butler AC, Roediger HL, 3rd. Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Med Educ*. 2013;**47**:674-682.
[26] Larsen DP, Butler AC, Roediger HL, 3rd. Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. *Med Educ*. 2009;**43**:1174-1181.
[27] Kromann CB, Jensen ML, Ringsted C. Test-enhanced learning may be a gender-related phenomenon explained by changes in cortisol level. *Med Educ*. 2011;**45**:192-199.
[28] Cantillon P. Do not adjust your set: the benefits and challenges of test-enhanced learning. *Med Educ*. 2008;**42**:954-956.
[29] Greenhalgh T. Computer assisted learning in undergraduate medical education. *BMJ*. 2001;**322**:40-44.
[30] Cook DA. The failure of e-learning research to inform educational practice, and what we can do about it. *Med Teach*. 2009;**31**:158-162.

Figure legends

Figure 1: Flow of participants through the study. black boxes, testing cases; grey boxes, studying only cases. Contamination occurred when students were erroneously exposed to the wrong presentation format (by reporting to the computer room assigned to the other group) at least once.

Figure 2: Learning outcome. Student performance in the entry and exit exam as well as the retention test for testing (black line) and studying items (grey line), respectively. Error bars indicate standard errors of the mean. ** $p<0.001$ in a paired t test comparing student performance in intervention and control items.

Figure 3: Performance trajectories. Proportion of intervention items answered correctly as a function of prior performance in identical items. Proportions of each category are connected by arrows from the preceding test.

Tables

Table 1: Teaching and e-seminar content. T, testing (cases contained key feature questions); S, studying (cases did not contain key feature questions).

Week	Content covered		E-seminar format	
	Formal teaching	**E-seminars**	**Group A**	**Group B**
1	Coronary artery disease	all (entry exam)	T	
2	Heart Failure	Coronary artery disease	T	S
3	Valvular disease	Heart failure	S	T
4	Arrhythmias	Valvular disease, coronary artery disease	T	S
5	Respiratory disease	Arrhythmias, heart failure	S	T
6	Peripheral artery disease, pulmonary em bolism, myocarditis, pericarditis	Respiratory disease, valvular disease, coronary artery disease	T	S
7	Nephrotic syndrome	Peripheral artery disease, pulmonary embolism, myocarditis, pericarditis, arrhythmias, heart failure	S	T
8	Electrolyte homeostasis	Nephrotic Syndrome, respiratory disease, valvular disease	T	S
9	Renal failure	Electrolyte homeostasis, peripheral artery disease, pulmonary embolism, myocarditis, pericarditis, arrhythmias	S	T
10	Anaemia	Renal failure, nephrotic syndrome, respiratory disease	T	S
11	Lymphoma	Anaemia, Electrolyte homeostasis, peripheral artery disease, pulmonary embolism, myocarditis, pericarditis	S	T
12	Solid tumours	all (exit exam)	T	

Table 2: Proportions of correct answers for the 30 key feature items in the exit exam and the retention test. * $p<0.05$ in a χ^2 test comparing re-testing and re-studying. AGB, arterial blood gases; ACE, angiotensin converting enzyme; CAD, coronary artery disease; CO_2, carbon dioxide; COPD, chronic obstructive pulmonary disease; CRB, confusion/respiratory rate/blood pressure; CT, computed tomography; ECG, electrocardiogram; FEV_1, forced expiratory volume in 1 second; PAD, peripheral artery disease ; PE, pulmonary embolism; VC, vital capacity

(see next page for the complete table)

Diseases	Key features	Exit exam		Retention test	
		Testing condition	Studying condition	Testing condition	Studying condition
Pulmonary embolism	Diagnosis of pulmonary embolism	59.5	48.9	57.1	64.4
	Wells Score before D-dimer testing	76.2	60.0	76.2	51.1*
	Thorax CT scan to confirm PE	42.9	55.6	26.2	44.4
	Right ventricular strain for risk stratification	45.2	22.2*	42.9	24.4
	Fibrinolysis for unstable pulmonary embolism	59.5	51.1	54.8	37.8
Arterial hypertension	Diagnosis of secondary hypertension	40.5	20.0*	35.7	24.4
	Diagnosis of diastolic dysfunction	21.4	20.0	23.8	17.8
	Diagnosis of ACE inhibitor cough	88.1	82.2	73.8	86.7
Hyponatraemia	Hospital admission for hyponatraemia	31.0	6.7*	35.7	26.7
	Thiazide diuretics as cause of hyponatraemia	78.6	60.0	59.5	60.0
	Diagnosis of central pontine myelinolysis	40.5	17.8*	31.0	15.6
Atrial fibrillation	Orthostatic challenge after syncope	54.8	46.7	61.9	55.6
	ECG diagnosis of tachyarrhythmia	66.7	66.7	64.3	62.2
	CHA_2DS_2-VASc score for anticoagulation	38.1	33.3	64.3	46.7
	Brain CT scan for suspected stroke	71.1	53.3	54.8	71.1
Lupus erythematodes	Diagnosis of Nephrotic Syndrome	75.6	59.5	62.2	57.1
	Diagnosis of systemic lupus erythematodes	71.1	59.5	75.6	64.3
	Renal biopsy to confirm lupus nephritis	71.1	59.5	71.1	69.0
	Immunosuppressive treatment for lupus nephritis	71.1	64.3	71.1	73.8
COPD	Diagnosis of COPD	66.7	76.2	75.6	71.4
	Confirmation of COPD by $FEV_1/VC<70\%$	68.9	35.7*	51.1	28.6*
	ABG analysis for suspected CO_2 intoxication	42.2	14.3*	48.9	21.4*
	Treatment of CO_2 intoxication by NIV	55.6	45.2	53.3	42.9
Pneumonia	Diagnosis of pneumonia in a chest X-ray	71.1	64.3	60.0	57.1
	CRB-65 score for hospital admission	53.3	28.6*	42.2	21.4*
Hyperthyroidism	Diagnosis of hyperthyroidism from lab results	86.7	71.4	82.2	81.0
	Stopping amiodarone in a pt. with hyperthyroidism	57.8	40.5	51.1	40.5
Pulmonary fibrosis	Diagnosis of pulmonary fibrosis	53.3	38.1	44.4	42.9
	Amiodarone as cause of pulmonary fibrosis	57.8	54.8	75.6	71.4
	Indication for long-term oxygen therapy	57.8	50.0	48.9	35.7